JOB SALARY PREDICTION

Devashish Thakur
Graduate Student
dthakur@cs.stonybrook.edu

Ayush Sengupta
Graduate Student
aysengupta@cs.stonybrook.edu

ABSTRACT
The aim of this project is to predict the salary of a Job by looking at the content of a job description, job location, job category and various other features associated with a job advertisement. The project is inspired from an online Kaggle competition program[10]. We try linear and non-linear decision boundary algorithms on the dataset and try to find out which one gives us the best results. We also try cluster analysis for classifying dataset. Later we use the same cluster analysis to do unsupervised learning and try to find interesting patterns within clusters of the same salary range and between clusters of different salary range.

Keywords
Classification, Regression, Supervised, Unsupervised, Machine Learning.

1. INTRODUCTION
Supervised learning has been one of the proficient methods used in Machine Learning to predict labels in un-labeled datasets. Any supervised algorithm needs two datasets – Training and Test dataset. Training dataset is used to model the parameters of the Machine Learning algorithm. Once trained we use this model to predict labels in the un-labeled Test dataset. Supervised learning can be divided into two parts – Classification and Regression. Classification is used where labels are from discrete buckets and number of these buckets is limited. The trained model labels each test dataset into one of these buckets. The accuracy of the model is calculated by finding the fraction of the test data set that was classified into wrong bucket.

In Regression problem we predict a continuous output for each row in the test dataset. The error is calculated as the absolute value of the average error across all the samples.

Classifier and Regressors can be either linear or non-linear. Linear algorithms try to separate labels by using a linear decision boundary whereas non-linear algorithms try to find non-linear decision boundaries in the dataset. Visualization of the decision boundaries is a difficult task if the dataset belongs to a high dimensional subspace.

In this project we try classification and regression algorithm on our datasets and use 5- fold cross validation to find the best parameter of our model.

The rest of the paper is as follows. Section 2 will talk about Feature Description and Preprocessing. Section 3 talks about Classification algorithm applied to the dataset. Section 4 will talk about the regression algorithms applied to the dataset. The section will also show how we used K-means clustering algorithm with Decision Tree regression as a classification + regression solution to the dataset. Section 5 will show the results of the different algorithms. Section 6 will talk about the Unsupervised learning done on the dataset. Section 7 talks about the method we tried that did not work. All the implementation is done in Python using Scikit-Learn toolkit.[9]

2. Feature Description and Preprocessing
The dataset has 7 features. - Category, Company, Contract Time, Full Description, Location and Title. The value that has to be predicted is the salary.

To preprocess we need to convert this string features into vector of numbers. We use Tf-Idf vectorizer[1] to convert each feature into vector of columns.

$$\mathrm{tf}(t,d) = 0.5 + \frac{0.5 \times \mathrm{f}(t,d)}{\max\{\mathrm{f}(w,d) : w \in d\}} \qquad \mathrm{idf}(t,D) = \log \frac{N}{|\{d \in D : t \in d\}|}$$

Fig 1 – Tf and Idf formulae [1]

Tf stands for Term frequency and Idf stands for Inverse Document Frequency. The product of the two terms gives Tf-Idf score of each word present across set of documents. Tf-Idf gives high score to words that appear in a document but are less frequent across other documents.

In our case documents are each tuple in the training dataset. We use threshold to pick top 'k' words for each feature and use that to construct our training data matrix.

We use a threshold of 10,000 for full description and 1000 for other features.

After preprocessing, the size of the training data set is 246,380 x 18232 and Test data set is 48954 x 18232

3. Classification

Prediction of salary is a regression problem. We try to convert this into a classification problem by breaking the salary into 3 buckets.

The below diagram shows the distribution of salary across tuples in the training dataset.

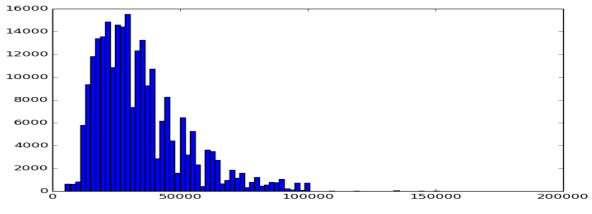

Fig 2 – Histogram of Salary distribution

The above plot does not look like a Normal curve and hence we cannot use this to bucket the dataset.

If we plot histogram of log(salary)

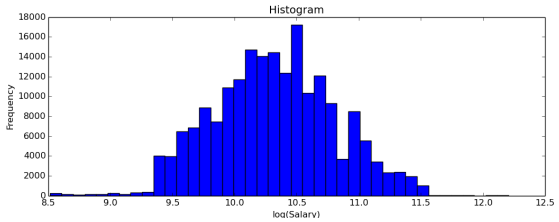

Fig 3 – Histogram of log(salary) distribution

We can see that in Fig-3 the distribution is closer to the normal distribution. We use log(salary) to divide the dataset into three buckets. Bucket 1 – **Mean – 1* SD < Salary < Mean + 1*SD** , Bucket 2 - **Mean + 1* SD < Salary** and Bucket 3 - **Mean – 1* SD > Salary**. We label the salary as 1,2 and 3 and use classification algorithms on the updated datasets.

We use 5-fold cross validation across all algorithms henceforth.

3.1 MultiNomial Naïve Bayes [3]

The first linear classifier we applied was MNB algorithm. We wanted to find if the dataset can be separated by a linear classifier.

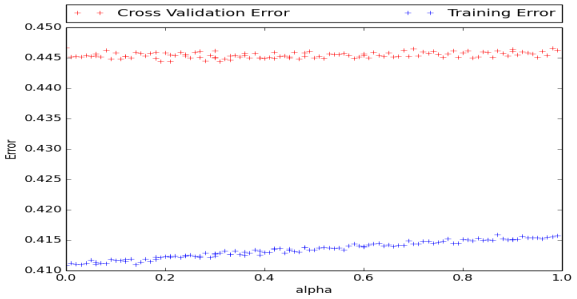

Fig 4 – MNB algorithm – Error Vs Alpha

Alpha is the regularization parameter used.

As we can see with increasing alpha, the cross validation and training error increased. Overall the best error was at alpha 0.445 which means 56% prediction accuracy.

3.2 Logistic Regression [4]

MNB assumes conditional independence of the features in the dataset and that may be the reason for the poor prediction accuracy. Hence we try Logistic regression since it doesn't make any conditional independence assumption about the dataset. This is also a linear classifier.

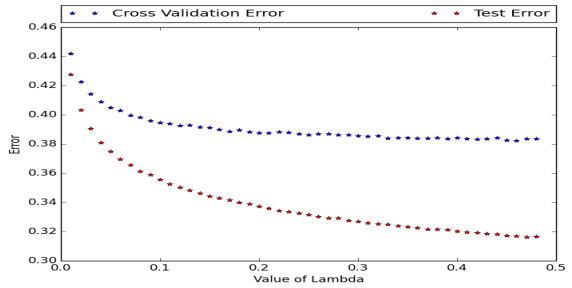

Fig -5 - LR algorithm – Error Vs Lambda

In Fig - 5, Lambda is the regularization parameter. We can see that the best accuracy that we get using LR is 60% on cross-validation dataset and 69% on training dataset. Neither MNB nor LR gave us good results. Hence we conclude that the dataset is not linearly separable. Henceforth we try our first non-linear classifier.

3.3 Decision Tree [5]

We apply Decision Tree algorithm on our data-set. We try to find the best height for our datasets.

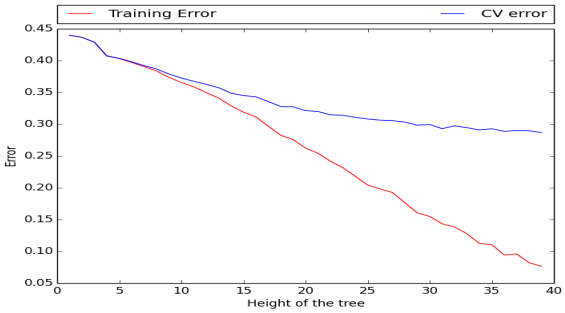

Fig 6 – Decision Tree - Error Vs Height of the Tree

This algorithm works nicely on our dataset. As we can see at height 38 we get 95% training data accuracy. At height 25 , we get predictive cross validation accuracy of 70% , which is best among all the classifiers we used till now.

Since Decision tree worked best on our dataset, we use it as our regression algorithm, which is discussed in the next section.

4. Regression

4.1 Decision Tree Regressor [5]

Since Decision Trees worked best for us as a classification algorithm, we decided to use it as a regression algorithm. The algorithm first splits the data set depending on the height given to the algorithm. Once the decision tree is constructed, each leaf is a bucket. Doing Linear Regression on each bucket gives the final output of salary.

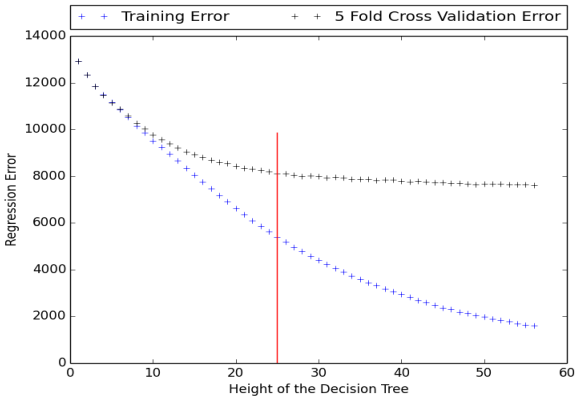

Fig – 8 – Decision Tree Regressor – Error Vs Height

As we can see in Fig 8, the ideal height is 25 which is almost same as what we got using Decision Tree Classification.

4.2 Random Forest Regressor [7]

This is an ensemble method that fits a number of classifying decision trees on various sub-samples of the dataset and use averaging to improve the predictive accuracy and control over-fitting. The height used is 25.

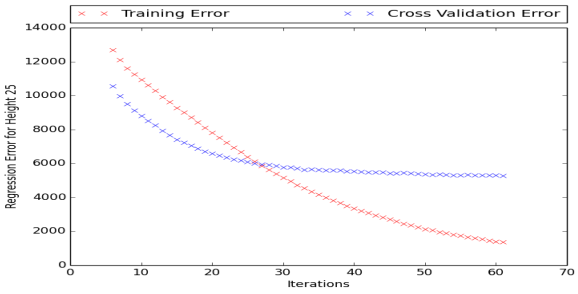

Fig 9 – Random Forest Regressor – Error Vs Iterations

We can see that at 40 iterations we get the least cross-validation error. And after 40, although the training error decreases exponentially , the cross-validation error almost remains the same.

4.3 K-Means Classifier [8]

Since our dataset was huge, many of the well known classification algorithms (SVD [11] etc) couldn't scale to the size of our dataset. Hence we tried another approach. We decided to cluster the dataset using K-means clustering algorithm. Once we cluster the data, we train Decision Tree Regressor for each cluster. When a test data point arrives, we find which cluster's centroid has the smallest Euclidean distance with this data point. Then we use the Decision Tree Regressor of that cluster to predict the salary of the data point.

The first task is to find the best K of the clustering algorithm.

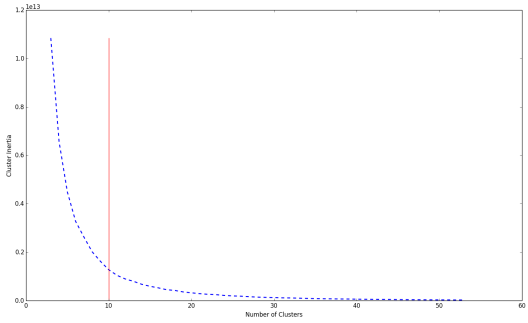

Fig 10 – K means – Cluster Inertia Vs Number of clusters

Cluster inertia is the sum of distances of samples to their closest cluster center. We see that the 'elbow point' in the curve is at K = 10 and hence we chose 10 as the number of clusters.

We then train Decision Tree Regressor for each cluster, the height is found using 5-fold cross validation.

The best height is 6 for cluster 5,6,8 and 7 for cluster 1,4,9 and 8 for cluster 2,3,7.

5. Results

We use two baselines to evaluate the results. The first baseline is provided in Kaggle. It is Random Forest which uses Decision Stump + 50 iterations + vector size of 100 using count vectorizer[2].

The second baseline is the Mean salary regressor. The Mean salary of the training dataset is 34174 and this regressor predicts this number for all the tuples in test dataset.

Algorithm	Test Error (Mean Error)
K Means + Decision Tree Regressor	**1134.33**
Random Forest (Height – 25 ,Iterations-40)	**1437.47**
Decision Tree Regressor (Height 25)	**2949.65**
Random Forest (Decision Stump + 50 iterations + Feature size – 100 + Count Vectorizer) - Baseline	**2363.37**
Mean Salary Regressor (34174) - Baseline	**9349.27**

Table 1 – Results of Regression Algorithms

As we can see in Table 1, K-Means + Decision Tree Regressor gives us the best results and it is almost half of the baseline error set by Kaggle. Random Forest with height 25 and 40 iterations also performs better than the Kaggle baseline.

6. Unsupervised Learning

We break the entire dataset into 10 clusters and then try to analyze the dataset in each cluster using K-means clustering. We also analyze common features across clusters. This gave us some important information about how job description changes across different salary range, which profession has the highest salary, and we also find words that are present in job description of multiple salary buckets but they have different meaning , and the meaning depends on the salary bracket.

We start with analysis of each cluster. Table 2 shows details of each cluster. If we look at the table , there are some interesting information. The website **cvjobstore.com** sends job advertisement only in the salary range 5k-18k. Similary **jobs.cabincrew.com** sends job requests only in the range 24k-31k.

Also if we see the contract type , we see that in higher packages , 82k – 200k , the number of contract type job is more than permanent jobs. It essentially means that people with more pays don't have permanent jobs.

Looking at the top titles across all the clusters we see that Doctors are the highest paid professional and Cleaners , administers, support workers are the least.

We also did inter cluster analysis. We plotted distribution of words in Job Title across different clusters.

Cluster Number	Salary Range	Job Type	Job Source	Top Titles
0	5k-18k	Permanent – 71% Contract – 28%	146 Unique - **cvjobstore.com**	Cleaner, Administrator, Support Worker, Account assistant, Credit Controller, Receptionist, Labourer
1	18k-24k	Permanent – 85% Contract – 15%	150 Unique - None	Recruitment Controller, Credit controller, Store Manager, Staff Nurse
2	24k-31k	Permanent – 86% Contract – 14%	155 Unique - **jobs.cabincrew.com**	Business Development Manager, Account Manager, Sales executive, Recruitment Consultant
3	31k-38k	Permanent – 87% Contract – 13%	153 Unique - None	Quantity Surveyor, Quality Engineer, Marketing Manager
4	38k-46k	Permanent – 90% Contract – 10%	151 Unique - None	Project Manager, Business Development Manager, Financial Consultant
5	46k-56k	Permanent – 87% Contract – 13%	146 Unique - None	Project Manager, Business Development Manager, HR Manager, HR Business Partner
6	56-68k	Permanent – 82% Contract – 18%	134 Unique - None	Financial Controller, Business Development Manager, Senior Product Manager
7	68k-82k	Permanent – 74% Contract – 26%	128 Unique - None	Financial Controller, Business Development Engineer, Finance director, Principal consultant , sales manager
8	82k-115k	Permanent – **59%** Contract – **41%**	119 Unique - None	Project Manager, Business Analyst, Finance Director, Financial Advisor
9	120k-200k	Permanent – **33%** Contract – **66%**	**34** Unique - None	Doctor: GP Locum in ***

Table 2 – Cluster Details of the 10 clusters

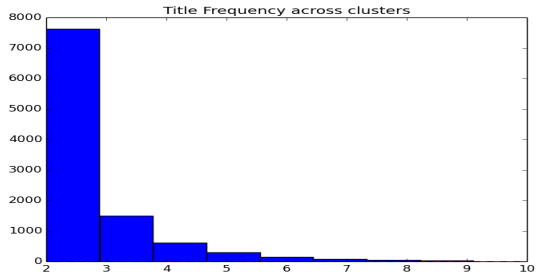

Fig -11 – Job Title Word Distribution

We can see in Fig – 11D, there are a lot of words which belong to 2 or 3 clusters , but there are few words that belong to all the clusters.

One of such word is **Project Manager.** The same title can be used in Job advertisements of multiple salary ranges. In case of lower salary (5k-18k), the Job title is 'Waking Night Project Manager' , whereas in clusters(82k- 200k) of higher salary range, the Job Title is 'Project Manager for private Banking London'.

Job Titles like '**Software/Firmware'** belong to only cluster 4 which is 38-46k , whereas Title like **'HealthCare/Education'** belong only to cluster 9. Titles like **'Broker/Technician'** belong only to cluster 3.

We also tried to analyze distribution of Job Locations across clusters.

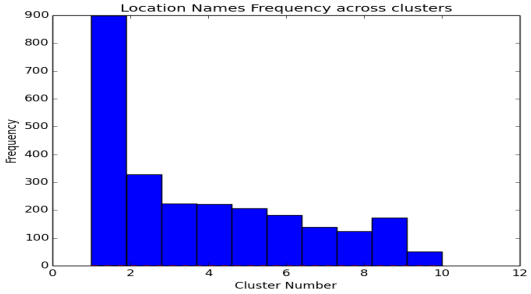

Fig – 12 – Job Location Distributions

Fig 12 shows the distributions of Job Locations across all salary ranges. There are a number of locations that belong to only one

cluster and very few locations that belong to all the clusters. Places like Howarh, Elvington and Eastgate are some of them that belong to only one cluster. Howarh and Elvington belong to cluster 0 , which means that job request for **Howarh** contains salary between 5k-18k. Another example is **Brightwell** which has job request only from range 31-38k (cluster 3).

Locations like Manchester , Suffolk , Yorkshire contains salary of all ranges.

We also tried to analyze the job description across clusters ,as to how job description changes across clusters and we found some interesting information. We first tried to find important word in each cluster using Tf-Idf. We then tried to see if there are words whose Tf-Idf score changes across clusters. Basically if Tf-Idf score is low in one cluster , it means its used a lot frequently across job description in that cluster and vice versa if the value of Tf-Idf is high.

We had a simple heuristic to find such outlier words. We analyzed all the words whose max – min > 2*avg , where max , min and avg is Tf-Idf scores of a word across clusters.

Word	Meaning
GP	Cluster 0 - "8k gp per month" , "paid on gp no threshold unless" Cluster 9 - "GP Unit Registered Nurse" The meaning changes from currency value to the word General Practitioner
Worksheets	Present in cluster 9 and 0 – Highest and Lowest paying jobs Highest – "Checking worksheets and auditing" Lowest – "Accurate completion of worksheets and creation of invoices" , "data entry in worksheets"
Notifications	Present in all clusters from 0 to 9 Highest Salary – "Analysis of notifications" ,"Designing of customer notifications", "follow notifications" Lowest Salary – "Push timely notification" , "Input and completion of Inspection notifications in SAP"
inspections	Cluster 0 – "cleaning inspections" , "Arranging property viewings inspections" Cluster 5 – "3rd party inspections companies" , "audit and inspection" Cluster 9 - "Undertake worksite inspections" , "health inspections"
Genius	Unique to cluster 5 and 6 , 46-68k salary "aspiring web genius" , "linguistic genius", "Photoshop genius"
Apodi	Brokerage company that offers salary between 18k-24k "Brokerage Support Officer opportunities in your area contact Ellie"
Vegetable	Found only in cluster 3 – salary 18—25k Related to catering and cooking jobs. All offers are from caterer.com.
nutitionist	Present only in cluster – 6 Salary 56-68k "Fertiliser Specialist, Crop nutitionist Crop"
Plastering	Unique to Cluster 0 and 1 (Low salary range) "small patch work plastering" , "include all aspect of plastering"

Table -3 – What word means what in each cluster

We can see how few word means different across clusters. A job description containing word '**GP**','**Notification**' or '**Inspection**' has different meaning , depending on the salary bucket of the Job. We also found words like '**Genius**' and '**Nutitionist**' which are unique to one cluster. We didn't plot histogram of distribution of Job description words across clusters as there were a lot of words and it was difficult to analyze outliers in that.

7. Other Methods Tried

Below are few methods we tried , but that didn't work on our dataset.

We have mentioned that Linear classifier MNB didn't work on our dataset.

We tried Adaboost [6] using MNB as weak classifier to find a non-linear classifier for our dataset.

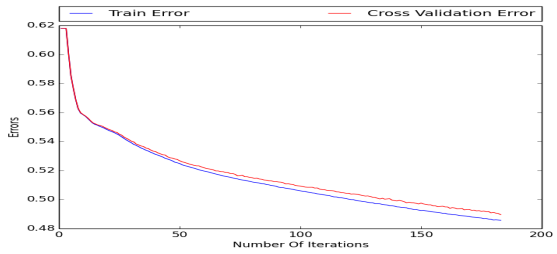

Fig – 12 – Adaboost with MNB

As you can see in Fig 12 , we tried till 180 iterations but the least Cross Validation error we get is of 0.48 , which means the best prediction value of 52%.

We also tried to apply PCA [12] on the original training dataset , in order to reduce the number of dimensions on our dataset but PCA didn't scale up to our dataset. We had the same problem with SVD [11] and hence we could not SVD classifier or regressor algorithm on our dataset.

8. Conclusion

In the above project we try to apply document classification techniques in a job advertisement dataset. We were able to get good results using K-Means classification + Decision Tree Regression and were able to do much better than the baseline model given by Kaggle.

We also did unsupervised learning on the dataset using K-means clustering and found some interesting pattern regarding distribution of salary across locations and Job Titles. We also found some interesting common words with different meaning across salary of different buckets.

As a future work we can apply more clustering algorithm to the dataset , like DBScan or Spectral clustering and we may be able to find some better/interesting patterns in the clusters of those algorithms.

9. References

[1] Scikit TfIdf Transformer http://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfTransformer.html

[2] Count Vectorizer - http://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html

[3] C.D. Manning, P. Raghavan and H. Schuetze (2008). Introduction to Information Retrieval. Cambridge University Press, pp. 234-265. http://nlp.stanford.edu/IR-book/html/htmledition/naive-bayes-text-classification-1.html

[4 Logistic Regression using]LIBLINEAR – A Library for Large Linear Classification http://www.csie.ntu.edu.tw/~cjlin/liblinear/

[5] L. Breiman, J. Friedman, R. Olshen, and C. Stone. Classification and Regression Trees. Wadsworth, Belmont, CA, 1984.

[6] Y. Freund, R. Schapire, "A Decision-Theoretic Generalization of on-Line Learning and an Application to Boosting", 1995

[7] L. Breiman, "Random Forests", Machine Learning, 45(1), 5-32, 2001.

[8] An Efficient k-Means Clustering Algorithm: Analysis and Implementation – Tapas Kanungo

[9] Scikit Learn toolkit - http://scikit-learn.org/

[10] Job Salary Prediction - https://www.kaggle.com/c/job-salary-prediction

[11] Wu, Lin and Weng, "Probability estimates for multi-class classification by pairwise coupling". JMLR 5:975-1005, 2004

[12] Principal Component Analysis http://support.sas.com/publishing/pubcat/chaps/55129.pdf